The Garden of Forking Paths in Visualization: A Design Space for Reliable Exploratory Visual Analytics

Position Paper

Xiaoying Pu* Matthew Kay†

University of Michigan, Ann Arbor

Abstract

Tukey emphasized decades ago that taking exploratory findings as confirmatory is "destructively foolish". We reframe recent conversations about the reliability of results from exploratory visual analytics—such as the multiple comparisons problem—in terms of Gelman and Loken's *garden of forking paths* to lay out a design space for addressing the forking paths problem in visual analytics. This design space encompasses existing approaches to address the forking paths problem (multiple comparison correction) as well as solutions that have not been applied to exploratory visual analytics (regularization). We also discuss how perceptual bias correction techniques may be used to correct biases induced in analysts' understanding of their data due to the forking paths problem, and outline how this problem can be cast as a threat to validity within Munzner's Nested Model of visualization design. Finally, we suggest paper review guidelines to encourage reviewers to consider the forking paths problem when evaluating future designs of visual analytics tools.

Index Terms: Human-centered computing—Visualization—Visualization design and evaluation methods

1 Introduction

> *In this garden of forking paths, whatever route you take seems predetermined, but that's because the choices are done implicitly.*
>
> — Gelman and Loken, 2013 [9]

The replication crisis in scientific fields like psychology and medicine is due in part to researchers' failure to distinguish between exploratory and confirmatory analyses. In theory, *confirmatory* analyses are those specified before looking at the data; any analyses that are not pre-specified which are carried out after an analysts see the data are *exploratory* [29]. This distinction is blurred in practice: A common practice in psychology, for example, has been to conduct many exploratory analyses and publish a small subset of them as if they were confirmatory [20]: *e.g.*, to hunt around in experimental results for significant p values, and then to publish only those analyses that are significant. Gelman and Loken liken this to wandering a *garden of forking paths* [9]: many paths through the garden (analyses) are tried, but only the final, successful path is reported. This practice has been known for years—*theoretically*—to lead to problems like overfitting [2], multiple comparison problems such as inflated false discovery rates [3], or publication bias (inflated effect sizes) [19]. Indeed, even as he introduced the term *exploratory data analysis* and advocated for its increased adoption, Tukey emphasized decades ago that taking exploratory findings as confirmatory is "destructively foolish" [28].

The recent replication crisis has been precipitated by empirical evidence that *theoretical* issues predicted by the forking paths problem have *actually appeared* in the literature. For example, the Open Science Collaboration conducted 100 replications of different psychology studies; the effect sizes in the replications showed "substantial decline" from the original studies—i.e., publication bias [25]. Well-known results in psychology have failed to be independently replicated—e.g., a meta-analysis of 32 independent, pre-registered replications of ego depletion found an effect size that is close to or equal to 0.[1] The once-theoretical warning that blurring the lines between exploratory and confirmatory analyses results in incorrect conclusions has become a practical reality.

We believe that the practice of wandering the garden of forking paths is typically neither malicious[2] nor even the analysts' fault: to our knowledge, data analysis tools rarely make any distinction between exploratory or confirmatory phases of analysis, nor any distinction between the reliability of conclusions drawn from either phase of analysis. It is necessary—as builders of visual analytics tools—that we confront how existing tool design has contributed to the development of the replication crisis and how better tool design may help to mitigate the crisis.

Indeed, the forking paths problem is not only present in visual analytics tools, it may even be made worse by them. Systems such as TimeSearcher [12] encourage users to rapidly explore different paths in the garden, slicing the dataset into possibly hundreds of different queries—representing different possible estimates or hypotheses—in a matter of seconds. Reda *et al.* [26], in a think-aloud study of an expert analyst's workflow with an interactive visualization tool, documented how the expert continually generated and tested hypotheses while exploring a real-world dataset. An analyst's freedom to explore can lead to high rates of false discoveries, as Zgraggen *et al.* showed with participants on synthetic data [33]. From a user-centered design perspective, it is not enough to simply build tools and hope that the user will handle those tools correctly. We must explicitly design visual analytics tools and workflows to support reliable inferences by mitigating the forking paths problem.

As the main contribution of this paper, we lay out a design space for solutions to the forking paths problem in visual analytics. The goal of our design space is to enable visual analytics tools to be constructed such that they produce more reliable findings. This design space encompasses existing approaches to addressing the forking paths problem—e.g., multiple comparison correction [33, 34]. To expand the design space, we draw a parallel between visual analytics and statistical modeling and machine learning and call attention to solutions that have not been explored, such as regularization or the incorporation of prior knowledge. We also draw a parallel

*e-mail: xpu@umich.edu

†e-mail:mjskay@umich.edu

[1] http://curatescience.org/collections/ego-depletion.html

[2] There is likely some small proportion of bad actors, but we do not believe they are the majority—this is why, like Gelman and Loken [9], we prefer not to use the term p-hacking.

to perceptual bias correction in visualization to suggest that such techniques may be useful in correcting biases induced in analysts' understanding of their data due to the forking paths problem. While we primarily focus on describing the design space, to help designers of visual analytics tools explicitly account for the forking paths problem, we also outline how it can be cast as a threat to validity within Munzner's Nested Model of visualization design [21], yielding suggestions for approaches to evaluating the extent of the problem in a new design. Finally, we suggest paper review guidelines to encourage reviewers to consider the forking paths problem when evaluating future designs of visual analytics tools.

2 Background

2.1 Making inferences and predictions in data exploration

Depending on their goal, an analyst might wish to perform two kinds of tasks in visual analytics systems: *non-generalization tasks* and *generalization tasks*.

Non-generalization tasks are those for which an analyst may wish to retrieve information from a dataset without generalizing findings to a population; a common example is performing a search. For example, a system like HomeFinder [31] helps the user find the ideal home in a given city. The system helps the user filter housing results based on their preferences. The user is unlikely to use it to make inferences about the broader "population" of houses—they simply want to find *particular* houses that match their preferences. Zhao et al. [34] viewed such visualizations as being designed for *descriptive statistics*.

Generalization tasks are those in which the analyst may want to make statistical generalizations from their data: for example, making inferences about a population or predictions about future results. This may include myriad systems that allow analysts to slice data by different variables to look for interesting patterns, particularly if the eventual goal is to fit a model to explain the data or to make predictions from it. Zhao, Zgraggen, and Kraska [33, 34] have framed such tasks in terms of Null Hypothesis Significance Testing (NHST), where the goal of a generalization task is to formulate and test hypotheses about a population. We attempt to adopt a wider view: that *generalization tasks* are any tasks in which the analyst wishes to develop or evaluate generalizable hypotheses, estimates, predictions, or understanding about a dataset, in any statistical framework (frequentist, Bayesian, or otherwise). This includes goals like making estimates about population parameters (instead of statistical tests), or making predictions about future data. All of these tasks are potentially vulnerable to the forking paths problem.

2.2 The forking paths problem in visual analytics

Based on the view that visual exploration of a dataset involves a continual process of formulating and testing hypotheses, the visualization community voiced concerns about the *multiple comparisons problem* in visual analytics systems [30, 33, 34]. From this perspective, to address the *multiple comparisons problem* is to correctly control false discovery rates. Zgraggen et al. [33] substantiated concerns about multiple comparisons in a study that allowed participants to explore synthetic datasets and generate insights without statistical testing, correction, or validation. Those exploratory insights, compared to the ground truth labels, had a low accuracy of 0.375 ± 0.297 and a high false discovery rate of 0.738 ± 0.296. It was only with confirmatory hypothesis testing and validation on a test set that the performance metrics improved. To obtain a roughly nominal FDR, Zgraggen *et al.* suggested to use the visual comparisons users made to correct generated insights after the exploration phase.

Hullman and Heer [14], in critiquing that work, suggested multiple alternative framings to the *multiple comparisons problem*, including a process of an analyst familiarizing themselves with the data, looking at the problem from a Bayesian perspective (allowing the incorporation of prior knowledge into the analysis process), or even a process where the analyst may just make observations to get familiarized with the data.

We define the *forking paths problem* as **unaddressed flexibility in data analysis that leads to unreliable conclusions**. This flexibility can any number of things, for example analyst decisions made contingent on data, unrestrained freedom to tune parameters in a model, or even the use of a flexible model prone to overfitting (e.g., ordinary least squares linear regression with the same number of parameters as observations [27]). In the NHST framework, the forking paths problem often manifests as a multiple comparison problem: the flexibility to conduct many statistical tests without accounting for the higher false discovery rate.

Our definition of the forking paths problem subsumes the multiple comparisons problem, accommodating more than just the frequentist NHST framework. For example, if an analyst were working in an estimation framework (such as Cumming's *New Statistics* [6] or Bayesian estimation [17]), the forking paths problem might manifest as an overfit model that leads to high out-of-sample RMSE. Such an analyst may care more about RMSE than FDR; indeed, using multiple comparison corrections would be inappropriate for this analyst's needs as multiple comparison corrections will not improve out-of-sample RMSE. Instead, regularization could be used (see Section 3.1.3).

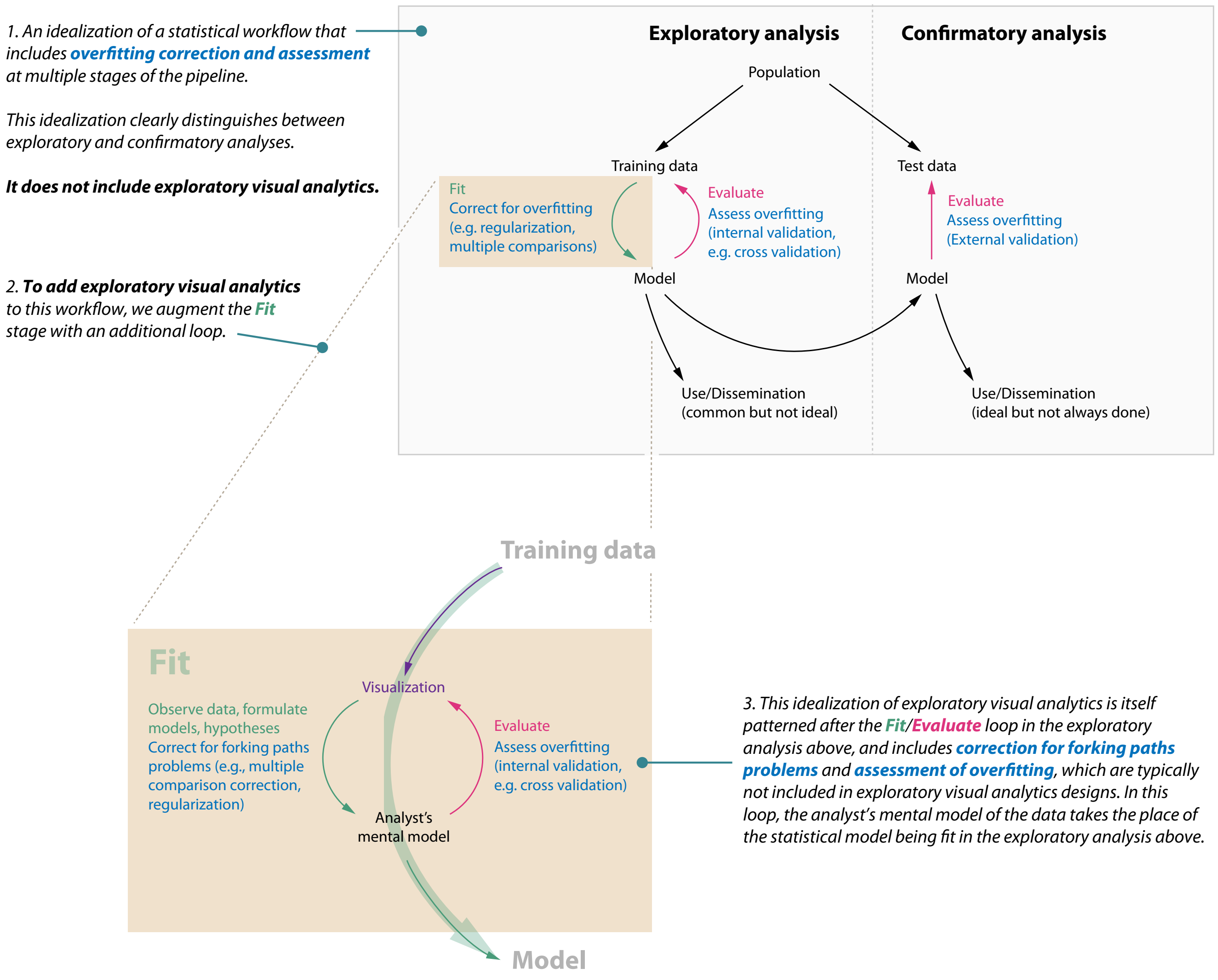
In some ways, the forking paths problem then becomes a question of how analysts update their mental model of the data as they explore it using a visualization tool; we liken this to the process of statistical model building and refinement in Figure 1.3. As one way to explain why analysts might be susceptible to the forking paths problem in visual analytics, we look to Wall *et al.*'s [30] metrics for measuring cognitive biases in visual analytics. These biases provide one reason that might cause analysts to be susceptible to the forking paths problem. For instance, the *oversensitivity to consistency* bias manifests as an analyst focusing only on data that seem to support their all-encompassing hypothesis and ignoring data that does not corroborate their hypothesis (even if present). With such biased exploration, the analyst is likely to overfit to their sliced data and generate exploratory conclusions that are unlikely to generalize. From this perspective, an analyst engaged in data exploration is building an explanation of the data (possibly hypotheses, or more generally some kind of *mental model* of the data), and cognitive biases may lead them to explore only subsets of the data, leading to biased inferences. In the end, the analyst considers only their final mental model of the data and the path that got them there, not other possible paths or models they might have considered along the way.

3 A forking paths-aware design space

We consider two aspects of how to design visual analytics systems to mitigate the forking paths problem: (1) how to assess and/or correct for the forking paths problem statistically, and (2) how to integrate such assessments or corrections into the visualization. These two aspects are drawn as the two dimensions in the design space shown in Figure 2. By explicating the approaches that previous research has used, we highlight areas of the design space that are currently underexplored, such as the use of regularization, or the incorporation of corrections directly into the visual presentation of the data.

3.1 Assessment/correction type

The vertical axis in Figure 2 shows the *assessment/correction type* dimension. This dimension is categorical, so the direction of it does not carry meaning. The vast majority of existing visualization systems have no correction for the forking paths problem. Recently, a small literature has developed around the multiple comparisons approach to addressing the problem. In this design dimension, we also want to propose the use of regularization (or relatedly, Bayesian



1. An idealization of a statistical workflow that includes ***overfitting correction and assessment*** *at multiple stages of the pipeline.*

This idealization clearly distinguishes between exploratory and confirmatory analyses.

It does not include exploratory visual analytics.

2. ***To add exploratory visual analytics*** *to this workflow, we augment the* ***Fit*** *stage with an additional loop.*

3. This idealization of exploratory visual analytics is itself patterned after the ***Fit/Evaluate*** *loop in the exploratory analysis above, and includes* ***correction for forking paths problems*** *and* ***assessment of overfitting****, which are typically not included in exploratory visual analytics designs. In this loop, the analyst's mental model of the data takes the place of the statistical model being fit in the exploratory analysis above.*

Figure 1: An illustration of how exploratory visual analytics might fit into a statistical modelling process, and how it might be improved via the incorporation of techniques to mitigate the forking paths problem by treating it as a statistical modelling process.

approaches) as alternatives that may be more applicable to some analysts' tasks.

3.1.1 No assessment or correction

If a visual analytics system does nothing to assess or correct for the forking paths problem, it falls into the *no assessment/correction* category. The vast majority of existing visual analytics systems fall into this category. An exemplary of this literature, TimeSearcher [12] allows an analyst to rapidly explore many different views of a set of time series without any statistical corrections from the system. This system is explicitly designed for generalization tasks—for example, helping stock brokers make predictions—thus it is vulnerable to the forking paths problem. Indeed, in the video describing the system (`https://youtu.be/VWx1TMcrb74`), the authors demonstrate rapid testing of many possible associations in an iterative, interactive exploration of the data—without any consideration of the forking paths problem—as an example of the use of the system. This is emblematic of the lack of consideration typically paid to this problem in the literature.

3.1.2 Multiple comparison corrections

One way of conceptualizing the garden of forking paths is as a multiple comparison problem. This is particularly appropriate if the goal of the analysis is ultimately hypothesis testing. From the Null Hypothesis Significance Testing (NHST) perspective, one goal in statistical analysis is to control the false discovery rate (the proportion of null hypotheses incorrectly rejected). By treating interactive visualization as a process of iterative hypothesis testing, Zgraggen, Zhao, and Kraska [33,34] account for the forking path problem using multiple comparison corrections. Zgraggen *et al.* [33] showed that a "mixing of exploratory and confirmatory testing" could achieve a similar false discovery rate (FDR) as validating on a test set. They treated the analyst's "implicit insights" during exploration as hypothesis testing and used the Benjamini and Hochberg multiple comparison correction procedure. Alternatively, Zhao *et al.* [34] use an approach based on α-investing, which keeps a budget to restrain the analyst's exploration and controls the marginal FDR (Figure 4-left).

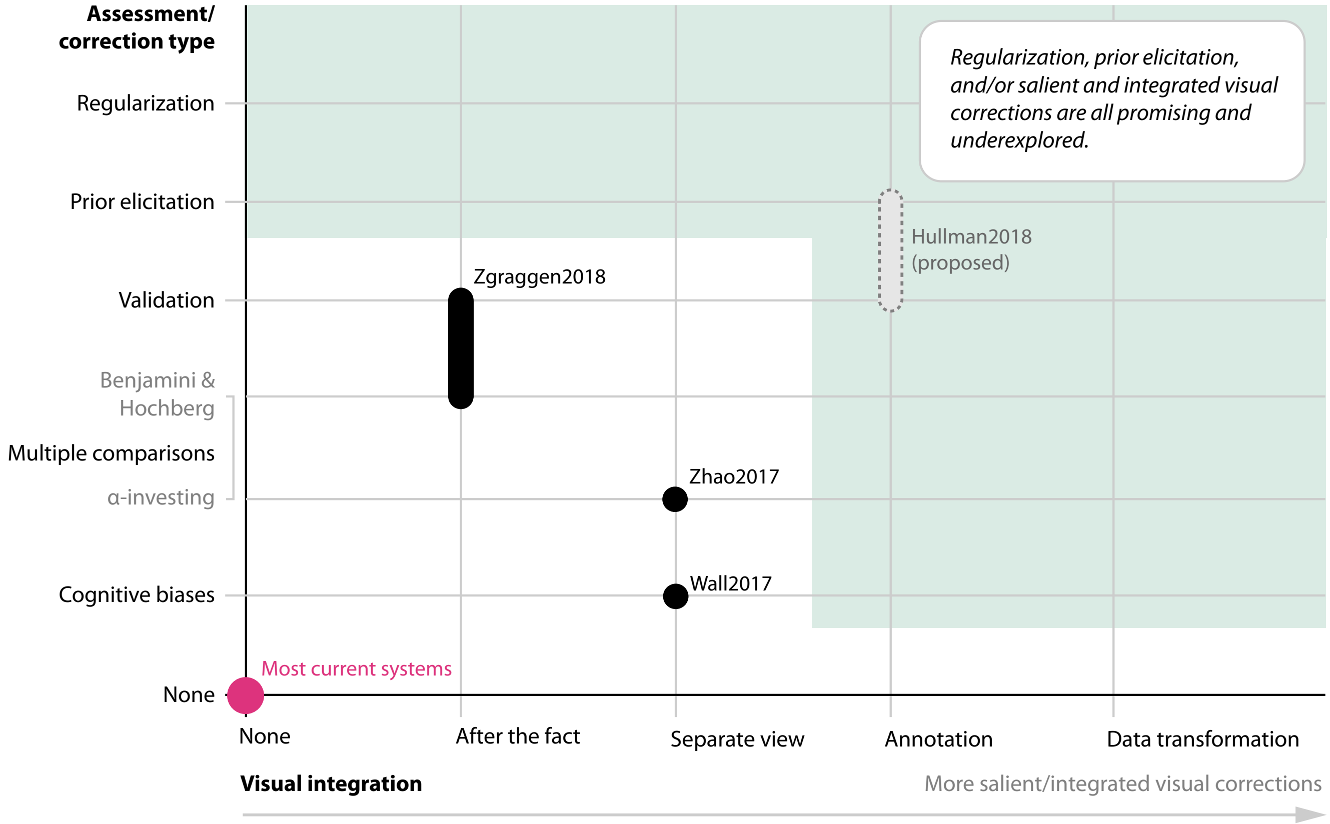


Figure 2: A design space for addressing the forking paths problem in exploratory visual analytics. The *assessment/correction type* axis is categorical, giving different ways of assessing or correcting for problems induced by the garden of forking paths. The *visual integration* axis indicates the level of salience of corrections within the visualization, or even (at the extreme) the presence of data transformations directly modifying raw data.

3.1.3 Regularization

Multiple comparison correction is appropriate if the analyst is working within a frequentist hypothesis testing framework, but may not match the analysts' goals if they are not interested in hypothesis testing. Analysts interested in estimation or prediction, for example, may be more interested in keeping estimation or prediction error low. Depending on the task or statistical framework, prediction errors can be measured using out-of-sample *root-mean squared error* or out-of-sample *KL-divergence* between the population distribution and the posterior predictive distribution [32], among others. An approach commonly adopted in statistics and machine learning for addressing the forking paths problem and reducing overfitting (and therefore improving out-of-sample error) is *regularization*. This use of regularization in visual analytics has not, to our knowledge, been explored.

Conceptually, one consequence of the forking paths problem is *overfitting*, meaning that the model or the analyst learns too much about incidental, *irregular* features of the sample at hand instead of learning about *regular* features that generalize to the population [18]. Thus, to *regularize* is to control the complexity of the model (or perhaps, the analysts' understanding of the data) in ways such that the model (the analysts' understanding) generalizes better.

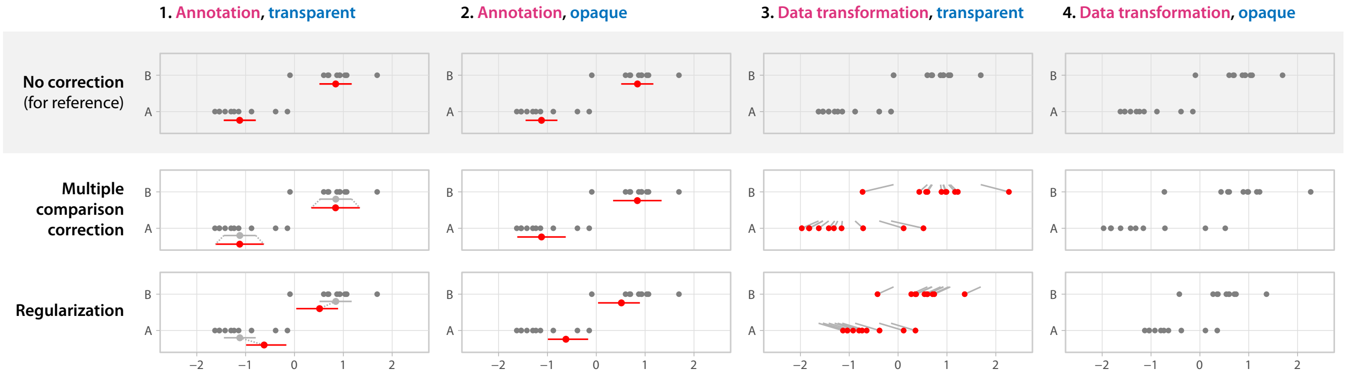
In this design dimension, we first review different regularization techniques in machine learning and statistics, and then we discuss how visual analytics can adopt regularization.

In Bayesian statistics, regularization can take the form of *regularizing priors*, also called *skeptical* or *weakly informed* priors [18]. Such priors are typically centered on some conservative (*a priori*) parameter value, e.g. 0 in the case of coefficients in a linear regression. Such a prior will shrink coefficients towards 0. The less data there is, the more coefficients are shrunk towards 0; in this way, irregular features that have little evidence for their presence are discounted in the model.

In a similar vein, Gelman *et al.* looked at the multiple comparison problem in terms of a hierarchical Bayesian model (also known as a random effects model), suggesting that from this perspective, we "usually don't have to worry about multiple comparisons" [8]. The partial pooling in hierarchical models is a form of regularization, shrinking estimates of different group means towards the global mean (see Figure 3). When the analysis task is estimation and not hypothesis testing, this approach is likely more applicable.

Regularization is also commonly used in machine learning to reduce overfitting. For instance, by adding a penalty term that depends on the characteristics of the parameters, regularization imposes model simplicity [23]. It is common in machine learning to discuss the resulting *bias-variance tradeoff* [11]: unbiased, unregularized estimates (such as those produced by ordinary least squares regression) will have higher out-of-sample error than regularized estimates (such as those produced by ridge regression [13]), which trade a little increase in bias for a greater decrease in variance, resulting in lower overall error [23].

Figure 1 shows the parallel we imagine between statistics/machine learning and an analyst's workflow during exploratory visualization, and how we imagine exploratory visualization fitting into a larger statistical workflow. We liken exploratory visual an-



*At the **annotation** level of the visual integration design dimension, we show corrected estimates and intervals alongside data. With **transparent** annotation, the relationship to uncorrected estimates is also be shown.*

*With **opaque** annotation, only the corrected estimates are annotated; the uncorrected estimates (and their relationship to corrected estimates) are not shown.*

*At the **data transformation** level of the visual integration design dimension, data are directly transformed so that uncorrected means and intervals calculated on the transformed data would be the same as corrected estimates on the original data. With **transparent** data transformation, the transformation is shown.*

*With **opaque** data transformation, only the transformed data are shown.*

Figure 3: Some possible combinations of different corrections for the forking paths problem with different annotation or data transformation approaches from our design space. Because only two groups are shown for space reasons, the effects of multiple comparison correction and regularization appear exaggerated in this example (the effect of those corrections applied to only two groups would be small).

alytics to exploratory data analysis, where rather than updating a statistical model the analyst is updating their mental model of the data. From this perspective, it is possible to imagine how statistical techniques—like regularization, multiple comparison correction, or validation—can fit into the visual analytics pipeline.

3.1.4 Eliciting priors

The Bayesian perspective offers a way to control regularization based on an analyst's prior knowledge. This is complementary to Hullman and Heer's call to solicit prior knowledge from analysts to help address the multiple comparison problem in visual analytics [14] (recast here as the forking paths problem). There is an extensive body of literature on expert prior elicitation [24]; recently-developed graphical approaches for eliciting priors from laypeople [10] might be particularly appropriate for integration into a visualization system.

3.1.5 Cognitive bias

Wall *et al.* [30] (Table 1 in that paper) outline several cognitive biases that analysts may be susceptible to when using visual analytics systems: the *vividness criterion* (e.g., relying more on specific/personal information than abstract information), *absence of evidence* (e.g., ignoring significant information that may be filtered from the current view), *oversensitivity to consistency* (e.g., ignoring data that does not support a broader model), *coping with evidence of uncertain accuracy* (e.g., fully accepting or rejecting evidence without considering for uncertainty), *persistence of impressions based on discredited evidence* (e.g., continuing to believe a hypothesis even after finding evidence against it).

Many of these biases dovetail with the forking paths problem: the *vividness criterion*, for example, might lead an analyst to pay undue attention to certain data, building a mental model that overfits to the particulars of a sample but generalizes poorly. Analysts may also be subject to problems of *persistence of impressions based on discredited evidence* if during some portion of exploratory analysis they come up with an interest—but incorrect—model or hypothesis, they may be slow to abandon it in the face of additional evidence later in exploration. Or worse, they may resist abandoning a model that was discovered during exploratory analysis but refuted during confirmatory analysis.

Wall *et al.* [30] describe a system that tracks how much an analyst interacts with different parts of a dataset and attempts to assess whether the analyst may be falling victim to some of the above biases. For example, they surface *data point coverage* and *data point distribution* metrics, which indicate if the analyst has interacted with an appropriately sized subset of the data (given how many interactions they have performed) and whether the distribution of those interactions are spread uniformly across the dataset (they also apply metrics of *coverage* and *distribution* to other aspects of a dataset, such as its attributes—i.e., variables). These metrics may help identify that an analyst has disproportionately explored one subset of the data, making them potentially subject to cognitive biases like the *vividness criterion* (if they place greater weight on data seen in more detail) or the *absence of evidence* bias (if they filter to a particular subset of the data and ignore the rest). Their system design surfaces biases to the user as linked views indicating the magnitude of each bias metric (data point coverage, data point distribution, etc), leaving what to do about it up to the analyst (Figure 4-right).

3.1.6 Validation

While approaches like regularization and multiple comparison correction are attempts to adjust models or estimates to *correct for* the forking paths problem, validation is one way to *assess* the extent to which it is a problem. Validation measures how well a statistical model generalizes—for example, by assessing how well the model is expected to perform on new data using metrics relevant to the analysts' estimation or prediction tasks (e.g., accuracy, precision, recall, root-mean-squared error, etc). Simply assessing model performance on a training set will lead to biased estimates of performance, because the model will have fit to some regular and irregular features of the sample at hand, but cannot distinguish between them. Depending on whether the model runs on a separate test set or not, the validation is considered either internal or external validation [15]. Internal validation, considered "a good first step" [15], can include cross-validation and bootstrap sampling. Using hold-out sets is a common way to do external validation.

In visual analytics, Zgraggen *et al.* [33] used a synthetic test set to validate insights that the participants discovered from exploration. For real-world data, a hold-out test set may not be easy to obtain.

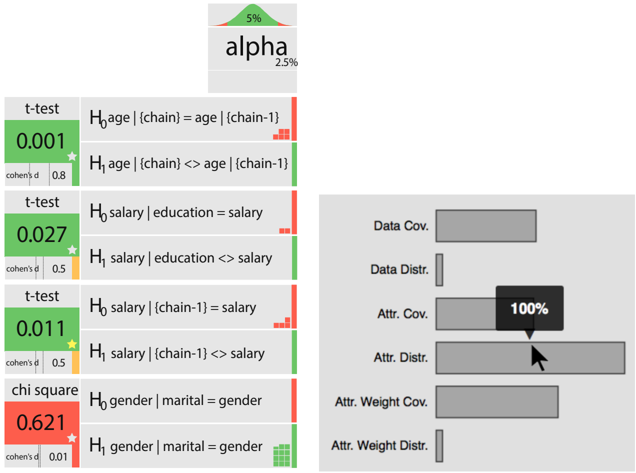

5%
alpha
2.5%
t-test
0.001
cohen's d
0.8
H_0 age | {chain} = age | {chain-1}
H_1 age | {chain} <> age | {chain-1}
t-test
0.027
cohen's d
0.5
H_0 salary | education = salary
H_1 salary | education <> salary
t-test
0.011
cohen's d
0.5
H_0 salary | {chain-1} = salary
H_1 salary | {chain-1} <> salary
chi square
0.621
cohen's d
0.01
H_0 gender | marital = gender
H_1 gender | marital = gender
Data Cov.
Data Distr.
Attr. Cov.
100%
Attr. Distr.
Attr. Weight Cov.
Attr. Weight Distr.

Figure 4: Two examples of *separate views* in visual integration of correction/assessment information. Both views are shown separately from the data view in each system. Left: the risk gauge simplified from Zhao *et al.* [34] that shows the analyst's α-investing budget and all hypotheses tested so far. Right: Wall *et al.*'s bar visualization shows assessment of possible biases in a user's interaction patterns that might suggest the user has not fully explored some aspects of the dataset [30].

Hence, one limitation to external validation is that this method "wastes" data that could have been used for model training.

One could imagine more explicitly incorporating the validation of an analyst's mental model of the data into the exploratory visual analytics process. For example, Hullman and Heer [14] propose that incorporating explicit user interfaces to elicit predictions from an analyst to help them refine and update their mental model of the data; Kim *et al.* [16] have found approaches like this to improve comprehension and recall of data. Perhaps such an approach could be adopted with hold-out sets and integrated directly into the exploratory analysis process.

3.2 Visual integration

In our design space in Figure 2, the x-axis indicates how integrated the assessment or correction of forking paths problems are into the visualization itself. This *visual integration* dimension is ordered by the saliency or extent of integration of the correction; from the least salient/integrated on the left to the most salient/integrated on the right:

1. **None**: The system does not display any information about forking path problems.

2. **Correction/assessment after the fact**: The system follows the analyst's interaction, assessing the problem during interaction without providing any online feedback. Any correction is done only after the end of exploration, *e.g.*, in a later model-fitting phase. For instance, one could imagine adapting an approach like that in Zgraggen *et al.*'s [33] study into a real analytics system by tracking users' interactions during exploratory analysis and then deriving a multiple comparison correction to be applied to models the users fit in a subsequent model-fitting phase.

3. **Separate view**: The interface displays correction information separate from the data visualization, e.g. as a view shown simultaneously (but not annotated directly on the visualization containing the data). Examples include Zhao *et al.*'s risk gauge [34] and Wall *et al.*'s [30] the bias metrics bar visualization. Figure 4 shows Zhao *et al.* and Wall *et al.*'s interfaces.

4. **Annotation**: The visualization of the data includes annotations that incorporate corrections for the forking paths problem. For example, for multiple comparisons corrections, confidence intervals become wider. For regularization, estimates may shrink closer to the global mean. (Figure 3.1, 3.2)

5. **Data transformation**: The data in the visualization is transformed to reflect the correction. Either in addition to or instead of using annotation to correct for forking paths problems, one could modify the data in the visualization such that an uncorrected inference on the visualized data is equivalent to a corrected inference on the original data. For example, one could shift sample data to have the same mean as a regularized estimate, or increase the variance of the data to have the same standard error implied by a multiple comparison-corrected estimate (Figure 3.3, 3.4). To our knowledge, such an approach has not been studied for application to the forking paths problem in visual analytics.

3.2.1 Transparency of annotation and data transformation

The *annotation* and *data transformation* categories may also vary in whether or not the correction is transparent to the user. With an **opaque visual correction**, the correction is applied but not clearly annotated as having been applied—e.g., a multiple comparison correction might be applied opaquely simply by widening the corresponding interval, and a data transformation by showing only the transformed data (Figure 3.2, 3.4). By contrast, a **transparent visual correction** would call attention to the fact that it has been applied—e.g., by showing both the small (uncorrected) and large (corrected) interval, or by showing the effect of a data transformation (Figure 3.1, 3.3).

3.2.2 Motivation for annotation and data visualization

We have several motivations for advocating for the use of *annotation* and/or *data transformation* to address the forking path problem. First, the use of direct annotation follows from well-established principles of effective visualization design, such as *Eyes Beat Memory* [22]: relying on users coordinating with views outside of the data (as in *separate view* designs) makes it less likely that they will take advantage of the provided assessments or corrections. Corrections should be visually salient and in close proximity to the relevant data in order to expect that they will be used.

Second, our proposal to explore data transformation approaches, while perhaps somewhat extreme and potentially controversial, draws inspiration from approaches used to correct perceptual biases. As Correll and Gleicher argue, since people have perceptual bias, the "correct" visualization may not always lead to "correct" knowledge or decisions [4]. To address perceptual biases, many techniques have been developed, from adjusting the area of symbols [7] to suppressing color range to reflect uncertainty [5]. Forking path *annotation* and *data transformation* corrections can act in a similar vein, modifying annotations and/or data to correct inference that may result from the exploration of the garden of forking paths.

Third, by putting these corrections front-and-center during the exploratory data analysis phase, we aim to prevent erroneous inferences before they are made: an analyst might make better decisions because the corrected visualizations stop them from discovering interesting (yet false) findings in the first place.

3.2.3 Feasibility of online annotation and data transformation

We believe that, in principle, a system can perform *annotation* and *transformation* online. For example, a system might track interactions in real time, and attempt to infer what aspects of the data the analyst is interested in (perhaps using Markov models, as in [30]).

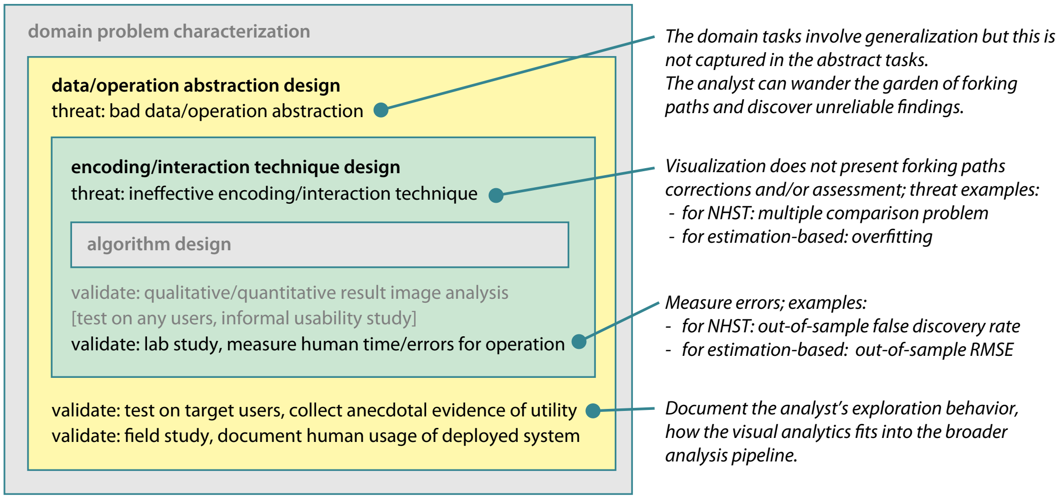

domain problem characterization
data/operation abstraction design
threat: bad data/operation abstraction
encoding/interaction technique design
threat: ineffective encoding/interaction technique
algorithm design
validate: qualitative/quantitative result image analysis
[test on any users, informal usability study]
validate: lab study, measure human time/errors for operation
validate: test on target users, collect anecdotal evidence of utility
validate: field study, document human usage of deployed system
The domain tasks involve generalization but this is not captured in the abstract tasks.
The analyst can wander the garden of forking paths and discover unreliable findings.
Visualization does not present forking paths corrections and/or assessment; threat examples:
- for NHST: multiple comparison problem
- for estimation-based: overfitting
Measure errors; examples:
- for NHST: out-of-sample false discovery rate
- for estimation-based: out-of-sample RMSE
Document the analyst's exploration behavior, how the visual analytics fits into the broader analysis pipeline.

Figure 5: Illustration of how The forking paths problem suggests threats to validity and possible evaluations at the second and third levels of Munzner's Nested Model of Visualization Design [21].

Since an existing design [34] already conducts corrections under the hood as the analyst interacts with data, we suggest the system reflect those corrections directly with *annotation* and *transformation* in real time.

We do anticipate that modifying annotations and transforming data can cause problems. These modifications could confuse the analyst or undermine the analyst's trust in the system. Suppose that the analyst filters to a subset of the data and performs a visual comparison. If the analytics system pools the data and annotation without explanation, the bewildered analyst might not know their analysis is being regularized. Even worse, without care in the design, an analyst might find that estimates change when they revisit them later if they have conducted more exploration in the interim—this might be avoided through approaches like α-investing [34], or through similar progressive approaches to correction like that outlined in our regularization example below. Some visualization approaches, such as transparent annotation (3.1, 3.3), might also lead to less confusion or mistrust than others. This is a complex problem that requires more work to develop clear guidance.

3.2.4 A hypothetical system with online correction and regularization

As one example of how a system *might* employ online correction and data transformation without changing data out from under the user, consider a hypothetical system employing data transformation alongside regularization. Such a system might only change data attributes the user has not previously seen—for example, the first time a user encounters a view with a given variable, it could be incorporated into a regularized model that corrects conditional means related to that variable that the user has not already seen. As the user explores more and more variables, each successive variable would be more and more regularized, while the first variables they looked at would be much less so. If the user is a domain expert, they might explore relationships they think are most fruitful *a priori* first; relationships they think less likely to be interesting—explored later—would be smoothed out more. This would implicitly help users combine expert knowledge (as expressed in how they interact with a system) with statistical corrections for data exploration. Meanwhile, the more and more the user explores paths that are contingent on the data (and not something they might have predicted *a priori*), the more and more the new data they look at is shrunken into uninteresting patterns.

4 Evaluating the Forking Paths Problem in Visual Analytics Systems

4.1 Forking paths problem as a threat to validity in the Nested Model

In visualization design and validation, we can view the forking paths problem as a threat to validity in Munzner's nested model [21] at two levels: the *data/operation abstraction design* level and the *encoding/interaction technique design* level. Figure 5 illustrates our use of the model.

At the data/operation abstraction design level, operations can refer to generic tasks such as "exposing uncertainty" and "concretizing relationships" [1, 21]. These generic tasks should yield reliable findings. If the domain tasks identified at the first level of the nested model involve some form of generalization, then the forking paths problem is a potential threat to validity. This can be assessed by asking if one of the abstract tasks the analyst wishes to do is related to generalization (example tasks might include hypothesis formation or testing, statistical estimation, model building, or prediction). If generalization tasks are involved in the users' workflow, a system design that allows the analyst to explore and to overfit to the data easily will not lead to reliable inferences for that task. Thus, there is a threat to the validity at the data/operation abstraction level. In other words, a system design at this level is valid only if it addresses the analyst's freedom to explore. This threat may be evaluated through observations of analyst behavior and workflow to document their exploration behavior and understand how it fits into the broader context of their analysis pipelines.

At the encoding/interaction technique design level, a typical threat is that a "chosen design is not effective at communicating the desired abstraction to the person using the system" [21]. In the case of the forking paths problem, this may occur if the visualization system does not effectively encode potential biases or uncertainty in data or estimates caused by the forking paths problem. This is particularly a problem if the workflow and abstract tasks of the user have no way of correcting for forking paths problems (such as overfitting or multiple comparisons) at a later stage of their analysis pipeline. In this case, designs that do not encode assessments or corrections for forking paths problems (e.g., through separate views, annotations, or data transformations) are ineffective at communicating the reliability of inferences derived from the data. This threat may be evaluated through lab studies that measure the impact of different designs on out-of-sample error relevant to the users' abstract tasks (e.g., FDR, root-mean-squared error, etc).

4.2 Review criteria

Stepping back, we need to assess how complicit we are in contributing to problematic results produced by visual analytics systems. A significant portion of the visualization field is concerned with the development of tools to help analysts make robust inferences from their data. If we publish—and endorse—interactive visualizations that are (1) intended to be used to make statistical generalizations (e.g., inferences, estimates, or predictions) and (2) do not account for the forking paths problem, we are effectively endorsing *p*-hacking machines.[3]

We would not expect statistics or machine learning journals to publish a new modeling technique if a straightforward application of it to the datasets it is designed for clearly leads to overfitting. Similarly, we should not expect the visualization community to publish visualization systems that, when combined with a user, task, and dataset it is designed for, also produces overfitted estimates. That is, swapping a model for a visualization + a user should not change how we evaluate the severity of the forking paths problem in an analysis pipeline.

In light of this, we suggest that assessment of the forking paths problem should become commonplace in the review of visualization tools, and should be reasonable grounds for revision (or even rejection) of a manuscript. The following questions may be helpful for a reviewer to ask:

1. Have the authors clearly articulated if any of the intended use cases for this system are generalization, inference, or prediction? If the system is not intended for any sort of statistical generalization, the forking paths problem is likely not relevant. For example, many search tasks would meet this criterion: a user searching for the ideal house [31] does not care to infer anything about the population of possible houses, they simply want to find a specific house that suits their needs.

2. If the system is intended for generalization, is it susceptible to the forking paths problem in its typical workflow? If it is, we believe the system should not be published. As we have described in this paper, there are many ways to potentially address this problem. One way the problem might be addressed is if the user and task analysis demonstrate that the tool is used in a workflow where the distinction between exploratory and confirmatory analysis is already clearly made (we expect this to be very rare). Another—possibly the best—approach to meet this criterion is for authors to integrate some method of addressing the forking paths problem into their design. If that is not done, we believe reviewers have just cause to reject a paper (per our explanation above). Another approach (acknowledging that until and unless such a criterion becomes a norm in the field, authors might be surprised to be judged based on it) could be to ask authors to include prominent warnings in their paper (e.g. in abstract and introduction) that the technique should not be deployed on generalization tasks unless first modified to account for the forking paths problem.

We should note that we do not consider criterion #2 to be met simply by asserting (without evidence) that analysts know the distinction between exploratory and confirmatory analysis and would account for this in their workflow. Given widespread problems with forking paths analysis in many fields, this assertion is dubious without evidence to the contrary from any particular field. Therefore, the more conservative approach is to directly guard against the forking paths problem in the design of any visual analytics tool designed for statistical generalization tasks.

5 Discussion

5.1 Designing to avoiding analysts' premature commitment

We highlight visual integration in our design space because we believe it may be crucial to effectively address the forking paths problem. As Wall *et al.* [30] note, people are vulnerable to the *persistence of impressions based on discredited evidence*: it may not be enough to track analysts' behavior and attempt to make statistical corrections for it after the fact; it would be better to reduce the likelihood that they draw erroneous conclusions in the first place. This motivates designs that track and correct for forking paths problems in real time through *annotation* or *data transformation*. The analyst that made an erroneous discovery during data exploration must later be convinced to forget it. The analyst that did not, already has.

5.2 Education

It behooves us to examine the implications of forking paths problems on how we teach visualization design. To our knowledge, when it comes to teaching principles for the construction of exploratory visualization tools, undergraduate and graduate level courses in information visualization tend to focus on technical implementation details and interactive visualization design patterns—both fundamentally important concepts; however, in our experience, concepts in statistics that are highly relevant to exploratory visualization (such as multiple comparison correction, regularization, Bayesian reasoning, or model validation) tend to be taught in other classes, and the implications and applications of these concepts to visualization are not made clear. We must consider how to integrate material on these concepts—and how they apply to reliable exploratory visualization—into existing information visualization curricula. Without that integration, we are relying on designers making those connections themselves, an ineffective strategy thus far.

6 Limitations and future work

As we have discussed in Section 3, a natural next step for addressing the forking paths problem is to fill in the gaps in our design space. Since some design options such as *data transformation* may be intrusive or confusing, evaluation of new designs should consider analysts subjective opinions in addition to performance metrics.

One limitation of our design space is that the *assessment/correction type* and the *visual integration* dimensions are not guaranteed to be exhaustive. As we bring awareness to the forking paths problem, we hope that the visualization community can help expand this design space.

7 Conclusion

This work was motivated by a reflection on how information visualization may have—in some small way, at least—helped precipitate the replication crisis, and how information visualization might help address the crisis. We believe that it is necessary to continue exploring the design space of possible solutions to the forking paths problem in visual analytics. We suggest that more statistical tools, such as regularization and validation, be brought to bear on the problem. These tools can be chosen in a principled way, depending on analysts' tasks (e.g., multiple comparison correction if tasks involve hypothesis testing and regularization if they involve estimation or prediction). By integrating these statistical tools and visual corrections into the analyst's system, we believe it is possible to prevent the forking paths problem from propagating further down an analysis pipeline. Finally, we urge the community to reconsider how we review exploratory visual analytics tools that are intended for generalization tasks: just as we would reject a visualization that leads to erroneous conclusions due to a poor encoding choice (3D pie charts!), so too should we reject a visualization that leads to erroneous conclusions due to a failure to consider the garden of forking paths.

[3] In the introductory footnote we said we did not like the term *p*-hacking. That ban was temporarily suspended here for effect.

References

[1] R. Amar and J. Stasko. BEST PAPER: A Knowledge Task-Based Framework for Design and Evaluation of Information Visualizations. In *IEEE Symposium on Information Visualization*, pp. 143–150, 2004. doi: 10.1109/INFVIS.2004.10

[2] M. Babyak. What You See May Not be What You Get: A Brief Introduction to Overfitting in Regression Type Models. *Psychosomatic Medicine*, 66(3):411–421, 2004. doi: 0033-3174/04/6603-0411

[3] Y. Benjamini and Y. Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.

[4] M. Correll and M. Gleicher. Bad for Data, Good for the Brain: Knowledge-First Axioms For Visualization Design. *Proceedings of the 1st Workshop on Dealing with Cognitive Biases in Visualisations*, 2014.

[5] M. Correll, D. Moritz, and J. Heer. Value-Suppressing Uncertainty Palettes. *Conference on Human Factors in Computing Systems - CHI '18*, 2018. doi: 10.1145/3173574.3174216

[6] G. Cumming. The New Statistics: Why and How. *Psychological Science*, 25(1):7–29, 2014. doi: 10.1177/0956797613504966

[7] J. J. Flannery. The relative effectiveness of some common graduated point symbols in the presentation of quantitative data. *Cartographica: The International Journal for Geographic Information and Geovisualization*, 8(2):96–109, dec 1971. doi: 10.3138/J647-1776-745H-3667

[8] A. Gelman, J. Hill, and M. Yajima. Why We (Usually) Don't Have to Worry About Multiple Comparisons. *Journal of Research on Educational Effectiveness*, 5(2):189–211, apr 2012. doi: 10.1080/19345747.2011.618213

[9] A. Gelman and E. Loken. The garden of forking paths: Why multiple comparisons can be a problem, even when there is no fishing expedition or p-hacking and the research hypothesis was posited ahead of time. *Department of Statistics, Columbia University*, 2013. doi: dx.doi.org/10.1037/a0037714

[10] D. G. Goldstein and D. Rothschild. Lay understanding of probability distributions. *Judgment and Decision Making*, 9(1):1–14, 2014.

[11] T. Hastie, R. Tibshirani, and J. H. Friedman. *The elements of statistical learning : data mining, inference, and prediction*. 2009.

[12] H. Hochheiser and B. Shneiderman. Dynamic Query Tools for Time Series Data Sets: Timebox Widgets for Interactive Exploration. *Information Visualization*, 3(1):1–18, mar 2004. doi: 10.1057/palgrave.ivs.9500061

[13] A. E. Hoerl and R. W. Kennard. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, 12(1):55, feb 1970. doi: 10.2307/1267351

[14] J. Hullman and J. Heer. Multiple Perspectives on the Multiple Comparisons Problem in Visual Analysis. 2018.

[15] A. C. J. Janssens and F. K. Martens. Prediction Research. Technical report, Atlanta, 2014.

[16] Y.-S. Kim, K. Reinecke, and J. Hullman. Explaining the gap: Visualizing one's predictions improves recall and comprehension of data. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pp. 1375–1386. ACM, 2017.

[17] J. K. Kruschke and T. M. Liddell. The Bayesian New Statistics: Two Historical Trends Converge. *SSRN Electronic Journal*, 2015. doi: 10.2139/ssrn.2606016

[18] R. McElreath. Overfitting, Regularization, and Information Criteria. In *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. CRC Press, 2016.

[19] P. E. Meehl. Why Summaries of Research on Psychological Theories are Often Uninterpretable. *Psychological Reports*, 66(1):195–244, feb 1990. doi: 10.2466/pr0.1990.66.1.195

[20] M. R. Munafò, B. A. Nosek, D. V. Bishop, K. S. Button, C. D. Chambers, N. Percie Du Sert, U. Simonsohn, E. J. Wagenmakers, J. J. Ware, and J. P. Ioannidis. A manifesto for reproducible science. *Nature Human Behaviour*, 1(1):1–9, 2017. doi: 10.1038/s41562-016-0021

[21] T. Munzner. A Nested Process Model for Visualization Design and Validation. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):921–928, 2009. doi: 10.1109/TVCG.2009.111

[22] T. Munzner. *Visualization Analysis and Design*. AK Peters Visualization Ser. A K Peters, Limited Taylor & Francis Group [Distributor], Natick : Florence, 2015.

[23] K. Murphy. *Machine learning: a probabilistic perspective*. The MIT Press, 2012.

[24] A. O'Hagan, C. E. Buck, A. Daneshkhah, J. R. Eiser, P. H. Garthwaite, D. J. Jenkinson, J. E. Oakley, and T. Rakow. *Uncertain Judgements: Eliciting Experts' Probabilities*. John Wiley & Sons, Ltd, Chichester, UK, jul 2006. doi: 10.1002/0470033312

[25] Open Science Collaboration. Estimating the reproducibility of psychological science. *Science*, 349(6251), aug 2015.

[26] K. Reda, A. E. Johnson, M. E. Papka, and J. Leigh. Modeling and evaluating user behavior in exploratory visual analysis. *Information Visualization*, 15(4):325–339, 2016. doi: 10.1177/1473871616638546

[27] M. K. Smith. Overfitting, 2014. `https://www.ma.utexas.edu/users/mks/statmistakes/ovefitting.html`.

[28] J. W. Tukey. *Exploratory data analysis*. Addison-Wesley series in behavioral science. Addison-Wesley Pub. Co., Reading, Mass., 1977.

[29] E.-J. Wagenmakers, R. Wetzels, D. Borsboom, H. L. J. van der Maas, and R. A. Kievit. An Agenda for Purely Confirmatory Research. *Perspectives on Psychological Science*, 7(6):632–638, 2012. doi: 10.1177/1745691612463078

[30] E. Wall, L. M. Blaha, L. Franklin, and A. Endert. Warning, Bias May Occur: A Proposed Approach to Detecting Cognitive Bias in Interactive Visual Analytics. *2017 IEEE Conference on Visual Analytics Science and Technology (VAST)*, 2017.

[31] C. Williamson and B. Shneiderman. The dynamic homefinder: Evaluating dynamic queries in a real-estate information exploration system. In *Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 338–346. ACM, 1992.

[32] Y. Yao, A. Vehtari, D. Simpson, A. Gelman, et al. Using stacking to average bayesian predictive distributions. *Bayesian Analysis*, 2018.

[33] E. Zgraggen, Z. Zhao, R. Zeleznik, and T. Kraska. Investigating the Effect of the Multiple Comparisons Problem in Visual Analysis. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*, pp. 1–12. ACM Press, New York, New York, USA, 2018. doi: 10.1145/3173574.3174053

[34] Z. Zhao, L. De Stefani, E. Zgraggen, C. Binnig, E. Upfal, and T. Kraska. Controlling False Discoveries During Interactive Data Exploration. *Proceedings of the 2017 ACM International Conference on Management of Data*, pp. 527–540, 2017. doi: 10.1145/3035918.3064019